Software Manual
Institute of Bioinformatics, Johannes Kepler University Linz

cn.FARMS: a latent variable model to detect copy number variations in microarray data with a low false discovery rate

— *Manual for the* `cn.farms` *package* —

Djork-Arné Clevert and Andreas Mitterecker

Institute of Bioinformatics, Johannes Kepler University Linz
Altenberger Str. 69, 4040 Linz, Austria
okko@clevert.de and mitterecker@bioinf.jku.at

Version 1.1.12, August 8, 2011

Contents

1 Introduction **3**

2 cn.FARMS: FARMS for CNV Detection **3**

3 Getting Started: cn.FARMS **5**
3.1 Quick start : Process SNP 6.0 array 5

4 Setup **11**

1 Introduction

The `cn.farms` package provides a novel copy number variation (CNV) detection method, called "cn.FARMS", which is based on our FARMS ("factor analysis for robust microarray summarization" (Hochreiter *et al.*, 2006)) algorithm. FARMS is since 2006 the leading summarization method of the international "affycomp" competition if sensitivity and specificity are considered simultaneously. We extended FARMS to cn.FARMS (Clevert *et al.*, 2011) for detecting CNVs by moving from mRNA copy numbers to DNA copy numbers.
In the following section we will briefly describe the algorithm and provide a quick start guide. For furhter information regarding the algorithm and its assessment see the `cn.farms` homepage at http://www.bioinf.jku.at/software/cnfarms/cnfarms.html.

2 cn.FARMS: FARMS for CNV Detection

cn.FARMS is described "in a nutshell" by the preprocessing pipeline depicted in Figure 1: **(1) Normalization** is performed at two levels. It has as *input* the raw probe intensity values and as *output* intensity values at chromosome locations which are leveled between arrays and are allele independent. At the *first level* normalization methods remove technical variations between arrays arising from differences in sample preparation or labeling, array production (e.g. batch effects), or scanning differences. The goal of the first level is to correct for array-wide effects. At the *second level* alleles are combined to one intensity value at a chromosome location and a correction for cross-hybridization between allele A and allele B probes is performed. Cross-hybridization arise due to close sequence similarity between the probes of different alleles, therefore a probe of one allele picks up a signal of the other allele. The optional corrections for differences in PCR yield can be performed at this step or after "single-locus modeling". We propose sparse overcomplete representation in the two-dimensional space of allele A and B intensity to correct for cross-hybridization

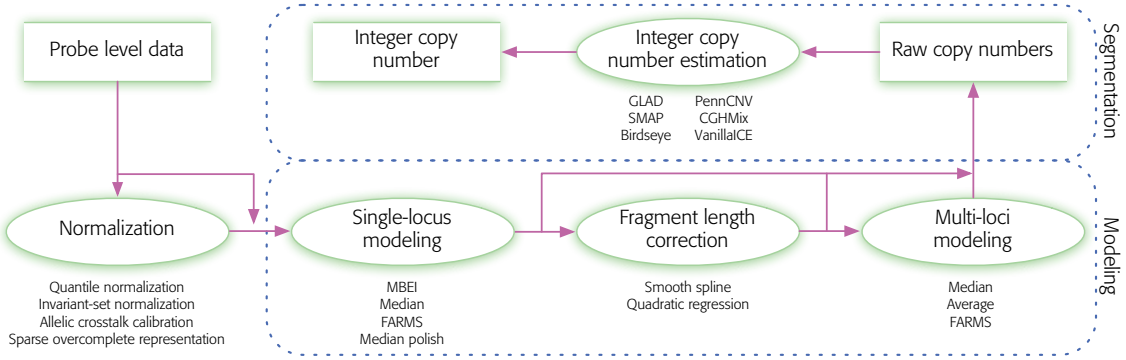

Figure 1: Copy number analysis for (Affymetrix) DNA genotyping arrays as a three-step pipeline: (1) Normalization, (2) Modeling, and (3) Segmentation. Modeling is divided into "single-locus modeling" and "multi-loci modeling" with "fragment length correction" as an optional intermediate step. The cn.FARMS pipeline is: normalization by sparse overcomplete representation, single-locus modeling by FARMS, fragment length correction, and multi-loci modeling by FARMS.

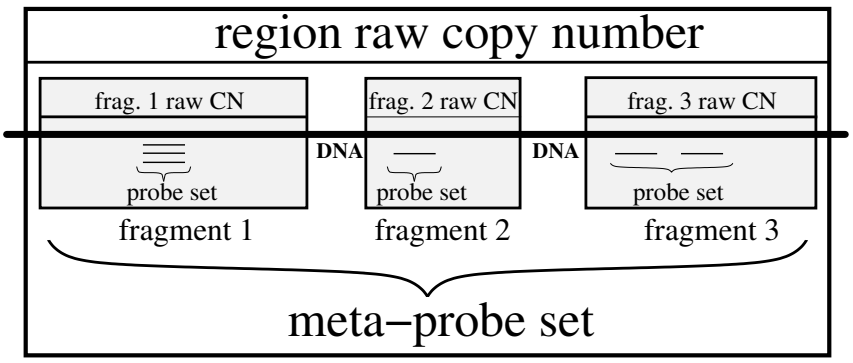

Figure 2: The copy number hierarchy probes-fragment-region. Fragment copy numbers serve as meta-probes used for "multi-loci modeling" which yields region copy numbers. Inner boxes: The probes which target a fragment (often at a SNP position) are single-locus summarized to a raw copy number of this fragment. Note, that instead of fragments a DNA probe loci can be summarized. Outer box: The raw fragment copy numbers are the meta-probes for a DNA region and are multi-loci summarized to a raw region copy number.

between allele A and allele B probes. Therefore we do not only estimate the AA and the BB cross-hybridization like CRMA (Bengtsson *et al.*, 2008) but also the AB cross-hybridization. The latter takes into account that hybridization and cross-hybridization may be different for the AB genotype, where for both allele probes target fragments are available and compete for hybridization. After allele correction, we follow CRMA and normalize by scaling the probes to a pre-specified mean intensity value. CNV probes which have only one allele are scaled in the same way. **(2) Modeling** is also performed at two levels. The *input* is the probe intensity values which independently measure the copy number of a specific target fragment or DNA probe locus. The *output* is an estimate for the region copy number. At the *first level*, "single-locus modeling" the probes which measure the same fragment are combined to a raw fragment copy number ("raw" means that the copy number is still a continuous values) — see Figure 2. These raw fragment copy numbers are estimated by FARMS. The original FARMS was designed to summarize probes which target the same mRNA. This can readily transfered to CNV analysis where FARMS now summarizes probes which target the same DNA fragment. Either both strands can be summarized together or separately where our default is the former. Nannya *et al.* (2005) suggested considering fragment characteristics like sequence patterns and the length because they affect PCR amplification. For example, PCR is usually less efficient for longer fragments, which lead to fewer copies to hybridize and result in weaker probe intensities. Following these suggestions cn.FARMS performs an optional intermediate level to correct for the fragment length and sequence features to make raw fragment copy numbers comparable along the chromosome. At the *second level*, "multi-loci modeling", the raw copy numbers of neighboring fragments or neighboring DNA probe loci are combined to a "meta-probe set" which targets a DNA region. The raw fragment copy numbers from single-locus modeling are now themselves probes for a DNA region as depicted in Figure 2. Again we use FARMS to summarize meta-probes and to estimate a raw copy number for the region. This modeling across samples is novel as previous methods only model along the chromosome. Multi-loci modeling considerably reduces the false discovery rates, because raw copy numbers of neighboring fragments or neighboring DNA probe loci must agree to each other on

the copy number, which reduces the likelihood of a discovery by chance. However, low FDR is traded against high resolution by the window size for multi-loci modeling, i.e. by how many raw copy numbers of neighboring fragments or neighboring DNA probe loci are combined. The more loci are combined, the more the FDR is reduced, because more meta-probes must mutually agree on the region's copy number. The window size for multi-loci modeling is a hyperparameter which trades off low FDR against high resolution. We recommend a window size of 5 as default, 3 for high resolution, and 10 for low FDR. Alternatively to a fixed number of CNV or SNP sites, the cn.FARMS software allows defining a window in terms of base pairs. In this case, multi-loci modeling may use a different number of meta-probes at different DNA locations, in particular for less than two meta-probes multi-loci modeling is skipped. Note, however that controlling the FDR is more difficult because a minimal number of meta-probes cannot be assured for each window and modeling with few meta-probes is prone to false discoveries. FARMS supplies an informative/non-informative (I/NI) call (Talloen *et al.*, 2007, 2010) which is used to detect CNVs. Additionally, the I/NI value gives the signal-to-noise-ratio of the estimated raw copy number. **(3) Segmentation** can afterwards be performed by `DNAcopy`.

3 Getting Started: cn.FARMS

As usual, it is necessary to load the `cn.farms` package:

```
library(cn.farms)
```

3.1 Quick start : Process SNP 6.0 array

The `hapmapsnp6` package is loaded for testing purpose.

```
> library("hapmapsnp6")
> celDir <- system.file("celFiles", package="hapmapsnp6")
> filenames <- dir(path=celDir, full.names=TRUE)
```

Next, the user specifies a working directory on the harddisk whereto save the results.

```
> workDir <- tempdir()
> dir.create(workDir, showWarnings=F, recursive=T)
> setwd(workDir)
```

For reasons of computational time and memory consumption `cn.farms` supports high-performance computation. The parameter `cores` specifies number of CPUs requested for the cluster and the parameter `runtype` indicates how the data matrix should be stored. `runtype = "ff"` creates a transient flat-file which will not be saved automatically. Whereas `runtype = "bm"` creates a persistent flat-file which can be save permanently.

```
> cores <- 2
> runtype <- "ff"
```

Next, the user specifies a subdirectory whereto save the flat-files.

```
> dir.create("ffObjects/ff", showWarnings=F, recursive=T)
> oligoClasses::ldPath(file.path(getwd(), "ffObjects"))
> options(fftempdir = file.path(oligoClasses::ldPath(), "ff"))
```

The directory (`celDir = "where/are/my/cel-files"`) which contain the cel-files has to be specified.

```
> celDir <- system.file("celFiles", package="hapmapsnp6")
> filenames <- dir(path=celDir, full.names=TRUE)
```

The following step will create the annotation file.

```
> if(exists("annotDir")) {
>        createAnnotation(filenames=filenames, annotDir=annotDir)
> } else {
>         createAnnotation(filenames=filenames)
> }
```

Afterwards, the data will be corrected for cross-hybridization and normalized.

```
> normMethod <- "SOR"
```

```
> ## normalization of SNP data
> if(exists("annotDir")) {
>         normData <- normalizeCels(filenames, method=normMethod, cores, alleles=T,
>                         annotDir=annotDir, runtype=runtype)
> } else {
>         normData <- normalizeCels(filenames, method=normMethod, cores, alleles=T,
>                         runtype=runtype)
> }
```

Now, the normalized data will be summarized at DNA probe loci. `summaryMethod <- "Variational"` indicates which FARMS approach should be used and `summaryParam$cyc <- c(10, 10)` specifies the number of iterations of the EM-algorithm. The parameter `summaryWindow` indicates whether DNA probe loci on the same DNA fragments are summarized together (`summaryWindow="fragment"`) or if the DNA probe loci are summarized separately (`summaryWindow="std"` is the default setting).

```
> summaryMethod <- "Variational"
> summaryParam <- list()
> summaryParam$cyc <- c(10)
> callParam <- list(cores = cores, runtype = runtype)
> slData <- slSummarization(normData, summaryMethod = summaryMethod,
+     summaryParam = summaryParam, callParam = callParam, summaryWindow = "std")
```

```
2011-08-08 16:46:21 |   Starting summarization
2011-08-08 16:46:21 |   Computations will take some time, please be patient
2011-08-08 16:46:21 |   Summarizing batch 1 ...
2011-08-08 16:46:22 |   Summarization done
Time difference of 0.7180002 secs
```

```
> show(slData)
```

```
ExpressionSet (storageMode: list)
assayData: 34 features, 268 samples
  element names: intensity, L_z, IC, lapla
protocolData: none
phenoData
  rowNames: NA10846 NA12146 ... NA19238 (268 total)
  varLabels: filenames gender
  varMetadata: labelDescription
featureData
  featureNames: 532152 532153 ... 745206 (34 total)
  fvarLabels: chrom start ... fragment_length2 (10 total)
  fvarMetadata: labelDescription
experimentData: use 'experimentData(object)'
Annotation: pd.genomewidesnp.6
```

```
> assayData(slData)$intensity[1:10, 1:5]
```

```
           [,1]      [,2]      [,3]      [,4]      [,5]
 [1,] 10.342002 10.803840 10.887975 10.755269 11.078718
 [2,]  9.163161  9.996646  9.981950  9.850188  9.457437
 [3,] 10.135801 11.210505 10.767266 10.908054 11.130400
 [4,]  9.033439  9.977772  9.458606  9.710699  9.713734
 [5,] 12.300862 12.224501 11.972838 12.146508 12.440324
 [6,] 11.704262 11.692252 11.557435 11.985427 11.662011
 [7,] 11.302287 11.120574 11.205731 10.787484 10.773983
 [8,]  9.507218  9.781986 10.217397  9.796442  9.777402
 [9,] 11.588298 11.400765 11.714276 11.907278 11.669675
[10,] 11.447309 11.032692 11.715353 11.522913 11.242266
```

```
> assayData(slData)$L_z[1:10, 1:5]
```

```
             [,1]         [,2]        [,3]         [,4]         [,5]
 [1,] -0.68146434 -0.219627213 -0.13549212 -0.268197864  0.055251714
 [2,] -0.52203358  0.311451683  0.29675558  0.164992996 -0.227757886
 [3,] -1.03323856  0.041465036 -0.40177357 -0.260985394 -0.038639342
 [4,] -0.67243501  0.271898813 -0.24726799  0.004825231  0.007860522
 [5,]  0.06263520 -0.013725371 -0.26538810 -0.091718464  0.202097403
```

```
 [6,]  0.07681114  0.064800299 -0.07001661  0.357976062  0.034559524
 [7,]  0.17473956 -0.006973599  0.07818358 -0.340063735 -0.353564639
 [8,] -0.25126123  0.023506863  0.45891787  0.037962312  0.018922187
 [9,]  0.03429350 -0.153239560  0.16027101  0.353273054  0.115670877
[10,]  0.01391892 -0.400698666  0.28196324  0.089522329 -0.191124222
```

Now, the intensity values of the non-polymorphic probes (CN-probes) were normalized.

```
> if (exists("annotDir")) {
        npData <- normalizeNpData(filenames, cores, annotDir=annotDir)
 } else {
        npData <- normalizeNpData(filenames, cores, runtype=runtype)
 }
```

This step combines non-polymorphic probes and single-locus summarized SNP-probes.

```
> combData <- combineData(slData, npData, runtype = runtype)
> show(combData)

ExpressionSet (storageMode: list)
assayData: 188 features, 268 samples
  element names: intensity
protocolData: none
phenoData
  rowNames: NA10846 NA12146 ... NA19238 (268 total)
  varLabels: filenames gender
  varMetadata: labelDescription
featureData
  featureNames: 70792 532141 ... 570013 (188 total)
  fvarLabels: chrom start end man_fsetid
  fvarMetadata: labelDescription
experimentData: use 'experimentData(object)'
Annotation: pd.genomewidesnp.6
```

In this final step intensity values of non-polymorphic probes and single-locus summarized SNP-probes are multi-locus summarized with a windows size of 5 probes (`windowParam$windowSize <- 5`). The window size for multi-loci modeling is a hyperparameter which trades off low FDR against high resolution. We recommend a window size of 5 as default, 3 for high resolution, and 7 for low FDR. Setting `windowParam$overlap <- TRUE` inidicates that the multi-locus summariaztion is done by step-wise moving the window over the genome. Alternatively to a fixed number of CNV or SNP sites, the cn.FARMS software allows defining a window in terms of base pairs. To make use of this option set `windowMethod <- "bps"`. In this case, multi-loci modeling may use a different number of meta-probes at different DNA locations, in particular for less than two meta-probes multi-loci modeling is skipped. Note, however that controlling the FDR is more difficult because a minimal number of meta-probes cannot be assured for each window and modeling with few meta-probes is prone to false discoveries.

```
> windowMethod <- "std"
> windowParam <- list()
> windowParam$windowSize <- 5
> windowParam$overlap <- TRUE
> summaryMethod <- "Variational"
> summaryParam <- list()
> summaryParam$cyc <- c(20)
> callParam <- list(cores = cores, runtype = runtype)
> mlData <- mlSummarization(slData, windowMethod = windowMethod,
+     windowParam = windowParam, summaryMethod = summaryMethod,
+     summaryParam = summaryParam, callParam = callParam)

2011-08-08 16:46:22 |   Starting summarization
2011-08-08 16:46:22 |   Computations will take some time, please be patient
2011-08-08 16:46:22 |   Summarizing batch 1 ...
2011-08-08 16:46:22 |   Summarization done

> names(assayData(mlData))

[1] "intensity" "L_z"       "IC"        "lapla"

> assayData(mlData)$intensity[1:10, 1:5]

          [,1]     [,2]     [,3]     [,4]     [,5]
 [1,] 10.51830 11.02014 10.91918 10.91897 11.01864
 [2,] 10.61055 11.37205 11.10092 11.18820 11.12730
 [3,] 11.22623 11.17878 11.12928 11.10760 11.17453
 [4,] 11.16531 11.12207 11.05674 11.06176 11.13546
 [5,] 11.56645 11.52526 11.45202 11.46987 11.54787
 [6,] 11.42628 11.35723 11.49604 11.53227 11.43893
 [7,] 11.34013 11.25488 11.43543 11.52618 11.38703
 [8,] 11.30726 11.21015 11.52576 11.61197 11.37157
 [9,] 11.36949 11.23962 11.68178 11.60176 11.41457
[10,] 11.34784 11.19053 11.75674 11.64901 11.40677

> assayData(mlData)$L_z[1:10, 1:5]

             [,1]          [,2]        [,3]        [,4]          [,5]
 [1,] -0.51098199 -0.0091432086 -0.11010793 -0.11030978 -1.064510e-02
 [2,] -0.56870234  0.1928008487 -0.07833413  0.00894956 -5.195163e-02
 [3,]  0.04697137 -0.0004764791 -0.04997542 -0.07165292 -4.726317e-03
 [4,]  0.04258345 -0.0006536843 -0.06598514 -0.06096191  1.273822e-02
 [5,]  0.04043301 -0.0007622544 -0.07399726 -0.05614699  2.185171e-02
 [6,]  0.01170655 -0.0573518778  0.08146067  0.11768897  2.434817e-02
 [7,]  0.01117620 -0.0740762497  0.10648134  0.19722565  5.807987e-02
```

```
 [8,] -0.02169338 -0.1188043505  0.19680905  0.28301784  4.262061e-02
 [9,] -0.04508354 -0.1749618879  0.26720034  0.18718662 -4.076551e-06
[10,] -0.06673986 -0.2240526240  0.34215749  0.23443277 -7.809875e-03
```

Next, the summarized data will be segmented in order to identify break points. Therefore we provide a parallelized version of `DNAcopy`.

```
> colnames(assayData(mlData)$L_z) <- sampleNames(mlData)
> segments <- dnaCopySf(x = assayData(mlData)$L_z[, 1:10], chrom = featureData(mlData)@data
+     maploc = featureData(mlData)@data$start, cores = cores, smoothing = FALSE)

Analyzing: Sample.1
Analyzing: Sample.1
Analyzing: Sample.1
Analyzing: Sample.1
Analyzing: Sample.1
Analyzing: Sample.1
Analyzing: Sample.1
Analyzing: Sample.1
Analyzing: Sample.1
Analyzing: Sample.1
Time difference of 0.3130002 secs

> head(featureData(segments)@data)

  chrom    start      end num.mark seg.mean individual
1    21 23655900 31642387        7  -0.1324    NA10846
2    23 47882970 47893260        9  -0.0594    NA10846
3    23 47895066 47942931       14  -0.1885    NA10846
4    21 23655900 31642387        7   0.0072    NA12146
5    23 47882970 47942931       23  -0.1209    NA12146
6    21 23655900 31642387        7  -0.0272    NA12239
```

To get further information, e.g. how to process Affymetrix 500K arrays, please check the followings demos.

```
> demo(package = "cn.farms")

Demos in package 'cn.farms':

demo01Snp6            Demo for an Affymetrix SNP6 data set
demo02Snp5            Demo for an Affymetrix SNP5 data set
demo03Snp500k         Demo for an Affymetrix 500K data set
demo04Snp250k         Demo for an Affymetrix 250K NSP data set
demo05Testing         Run the examples
```

The most recent `cn.farms` version can be found at http://www.bioinf.jku.at/software/cnfarms/cnfarms.html.

4 Setup

This vignette was built on:

```
> sessionInfo()

R version 2.13.1 (2011-07-08)
Platform: i386-pc-mingw32/i386 (32-bit)

locale:
[1] LC_COLLATE=C
[2] LC_CTYPE=English_United States.1252
[3] LC_MONETARY=English_United States.1252
[4] LC_NUMERIC=C
[5] LC_TIME=English_United States.1252

attached base packages:
[1] tools     stats     graphics  grDevices utils     datasets  methods
[8] base

other attached packages:
[1] DNAcopy_1.26.0      cn.farms_1.1.12     snowfall_1.84
[4] snow_0.3-6          oligoClasses_1.14.0 ff_2.2-3
[7] bit_1.1-7           Biobase_2.12.2

loaded via a namespace (and not attached):
 [1] Biostrings_2.20.2     DBI_0.2-5             IRanges_1.10.4
 [4] affxparser_1.24.0     affyio_1.20.0         grid_2.13.1
 [7] lattice_0.19-31       oligo_1.16.0          preprocessCore_1.14.0
[10] splines_2.13.1
```

References

Bengtsson, H., Irizarry, R., Carvalho, B., and Speed, T. P. (2008). Estimation and assessment of raw copy numbers at the single locus level. *Bioinformatics*, **24**(6), 759–767.

Clevert, D.-A., Mitterecker, A., Mayr, A., Klambauer, G., Tuefferd, M., Bondt, A. D., Talloen, W., Göhlmann, H., and Hochreiter, S. (2011). cn.FARMS: a latent variable model to detect copy number variations in microarray data with a low false discovery rate. *Nucleic Acids Research*.

Hochreiter, S., Clevert, D.-A., and Obermayer, K. (2006). A new summarization method for Affymetrix probe level data. *Bioinformatics*, **22**(8), 943–949.

Nannya, Y., Sanada, M., Nakazaki, K., Hosoya, N., Wang, L., Hangaishi, A., Kurokawa, M., Chiba, S., Bailey, D. K., Kennedy, G. C., *et al.* (2005). A robust algorithm for copy number detection using high-density oligonucleotide single nucleotide polymorphism genotyping arrays. *Cancer Research*, **65**(14), 6071–6079.

Talloen, W., Clevert, D.-A., Hochreiter, S., Amaratunga, D., Bijnens, L., Kass, S., and Göhlmann, H. W. H. (2007). I/NI-calls for the exclusion of non-informative genes: a highly effective feature filtering tool for microarray data. *Bioinformatics*, **23**(21), 2897–2902.

Talloen, W., Hochreiter, S., Bijnens, L., Kasim, A., Shkedy, Z., and Amaratunga, D. (2010). Filtering data from high-throughput experiments based on measurement reliability. *Proc. Natl. Acad. Sci. U S A*, **107**(46), 173–174.